# Compte-rendu du TP n°2 – K-means puis ACP

On n’arrive pas à en tirer quelques choses de la matrice de corrélation. C’est l’existence de plusieurs types qui peut expliquer cette matrice si dispersée. Il y a plusieurs types de populations ici donc c’est compliqué d’analyser la matrice de population sur l’ensemble de la population. Il faut donc appliquer K-means pour clusteriser le groupe et après pouvoir appliquer ACP et on verra des corrélations forcément plus forte/faible sur des composantes individuellement.

Faire attention pour Outcome car c’est des résultats (binaires) donc la matrice de corrélation entre une loi normale et binaire n’atteindra jamais 1. Donc prudence dans l’analyse des résultats.

Pour le nombre de dimensions à garder, 5 c’est pas mal. Mais cela dépend du nombre d’individus.

Cercle de corrélation CP1-2 : On remarque 3 groupes, dont 2 (insulino dépendant et lié à l’âge) + 1 qui participe aux 2 groupes.

CP\_1 = comportement moyen

CP\_2 = dispersion des individus.

CP3-4 : on comprend mieux sur les antécédents génétiques.

Dans le groupe insulino-dépendant : il y a 2 groupes, ceux liées aux antécédents, soit liée à autre chose (consommation, etc. = déclenchement externe)

CP5-6 : on commence à trop descendre dans le détail. C’est pas intéressant pour nous pour une analyse au global.

c.data=outcome

On superpose le cercle avec le graphique des individus.

* On observe un comportement plus dispersé avec les individus diabétiques (en jaune) car on réagit différemment en fonction de nos hygiènes de vie par rapport aux individus sains (avec le gros hama violet)
* Il n’y a pas 3 groupes ni 2 groupes, il y a 3 comportements avec un gradient (une évolution au fur et à mesure du temps de la glycémie par exemple)
  + On peut dire qu’il y a 2-3-4 groupes seulement lorsqu’on a étudié le graphique des individus avec l’ensemble des composants (ON A FAIT LE CHOIX DE…)

Avec un filtre sur le cluster 2 = les insulino dépendants, si on reproduit la matrice de corrélation, la variable « outcome » ne représente plus rien car elle est toujours égale à 1. Il faut faire attention à ne pas lire cette ligne. On constate cependant un lien entre âge et insuline (augmentation).

Avec K-means (10% du nombre de clusters par rapport au nb d’individus), on perd les valeurs extrêmes. Du coup l’ACP va être sur un comportement moyen.